Cognition 212 (2021) 104696

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Moral concerns are differentially observable in language

Brendan Kennedy [a,c,*], Mohammad Atari [b,c], Aida Mostafazadeh Davani [a,c], Joe Hoover [b,c], Ali Omrani [a,c], Jesse Graham [d], Morteza Dehghani [a,b,c]

[a] Department of Computer Science, University of Southern California, United States of America
[b] Department of Psychology, University of Southern California, United States of America
[c] Brain and Creativity Institute, University of Southern California, United States of America
[d] Department of Management, David Eccles School of Business, University of Utah, United States of America

ARTICLE INFO

Keywords:
Morality
Language
Text analysis
Moral foundations theory
Natural language processing

ABSTRACT

Language is a psychologically rich medium for human expression and communication. While language usage has been shown to be a window into various aspects of people's social worlds, including their personality traits and everyday environment, its correspondence to people's moral concerns has yet to be considered. Here, we examine the relationship between language usage and the moral concerns of Care, Fairness, Loyalty, Authority, and Purity as conceptualized by Moral Foundations Theory. We collected Facebook status updates ($N = 107{,}798$) from English-speaking participants ($n = 2691$) along with their responses on the Moral Foundations Questionnaire. Overall, results suggested that self-reported moral concerns may be traced in language usage, though the magnitude of this effect varied considerably among moral concerns. Across a diverse selection of Natural Language Processing methods, Fairness concerns were consistently least correlated with language usage whereas Purity concerns were found to be the most traceable. In exploratory follow-up analyses, each moral concern was found to be differentially related to distinct patterns of relational, emotional, and social language. Our results are the first to relate individual differences in moral concerns to language usage, and to uncover the signatures of moral concerns in language.

1. Introduction

Language is a fundamental medium for much of the human sciences; it is the "stuff of thought" (Pinker, 2007), the very material with which we communicate, express, teach, remember, and govern. Accordingly, it has been shown that the words people use convey rich information about their personality traits (Park et al., 2015; H. A. Schwartz et al., 2013), demographic characteristics (Pennebaker, Mehl, & Niederhoffer, 2003), and mental health (Rodriguez, Holleran, & Mehl, 2010), among other psychological factors (Tausczik & Pennebaker, 2010). Language is also a window into our moral worlds: through language, humans create and share religious ideas (i.e., in religious texts), conduct ideological debate (Clifford & Jerit, 2013), and express their personal values (Boyd et al., 2015). In fact, language, in facilitating communication and expression, is the backdrop of many measures of moral phenomena: psychologists compile moral linguistic stimuli (Clifford, Iyengar, Cabeza, & Sinnott-Armstrong, 2015; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014), design questionnaires on personal values and moral concerns (Graham et al., 2011; S. H. Schwartz et al., 2001), and conduct interviews on moral topics (Gilligan, 1977; Hallen, 2000; Kohlberg, 1981).

Motivated in part by the desire to investigate moral cognition in more ecologically valid ways (Hofmann, Wisneski, Brandt, & Skitka, 2014), over the past decade a growing body of research has used recorded language to chart the moral domain in "the wild" (Dehghani et al., 2016). In considering the interaction between political content and moral rhetoric online, Grover, Bayraktaroglu, Mark, and Rho (2019) investigated immigration policy debates in the United States on Twitter, finding that pro-immigration and anti-immigration tweets contained differing types of moral content. Similarly, Mooijman, Hoover, Lin, Ji, and Dehghani (2018) found, in a large-scale analysis of Twitter posts, that voicing moral concerns preceded the escalation of violence at a protest.

The sharing of moral language has also drawn attention, with Brady, Wills, Jost, Tucker, and Van Bavel (2017) finding that moral messages on Twitter were shared at a greater rate than others (cf. Burton, Cruz, &

* Corresponding author at: 362 S. McClintock Ave, Los Angeles, CA 90089-161, United States of America.
E-mail address: btkenned@usc.edu (B. Kennedy).

https://doi.org/10.1016/j.cognition.2021.104696
Received 15 May 2020; Received in revised form 1 March 2021; Accepted 19 March 2021
Available online 31 March 2021

Hahn, 2019). Similarly, recent work has examined the role of moral framing in political persuasion (Day, Fiske, Downing, & Trail, 2014; Feinberg & Willer, 2015; Voelkel & Feinberg, 2018). For example, Feinberg and Willer found that political arguments framed according to the audience's moral concerns were more successful. In all, these recent works on morality and language have highlighted moral rhetoric in political contexts as well as moral content posted on and shared via social media.

Despite the obvious ties between morality and language, as well as the recent interest in studying moral rhetoric and framing, moral language has yet to be investigated directly in relation to individual-level moral concerns. In previous work, moral language is used as a proxy to individuals' moral concerns; in other words, it is implicitly assumed that there is a one-to-one correspondence between composing moral content and being morally predisposed. Our aim in this work is to empirically test whether this association exists. Using a wide range of methods, we explore how moral concerns of Care, Fairness, Loyalty, Authority, and Purity as conceptualized by Moral Foundations Theory (MFT; Graham et al., 2013; Haidt & Joseph, 2004), and measured by the Moral Foundations Questionnaire (MFQ; Graham, Haidt, & Nosek, 2008), relate to the usage of moral language. Establishing a link between moral concerns and moral language would benefit the internal validity of ongoing observational research of moral language.

The interrelatedness of morality and language also points to another untested implicit assumption in the literature. The previous work in the domain of moral rhetoric makes the implicit assumption that the only association between moral concerns and language is in *moral* language. However, moral language is a narrow "slice" of language as a whole, and, as we have discussed, there is a rich supply of psychological information that can be found in peoples' language. In addition to explicit moral rhetoric, language offers the opportunity to "excavate people's thoughts, feelings, motivations, and connections with others" (Pennebaker, 2013, p. xi), generating insight into the psychological states of speakers in relation to their moral concerns. Furthermore, recent theoretical work has argued that language facilitates multiple "moral functions" from a social cognition standpoint (Li & Tomasello, 2021): humans initiate, preserve (i.e., maintain through generations, justify against alternatives), revise, and act on morality uniquely through language. As such, it is artificially limiting to suggest that morality occurs in language usage only when individuals use explicitly moral words; rather, it is likely that individuals take part in moral debates and topics by engaging in a wide array of social language. This perspective reinforces the importance of examining the relation between morality and language usage in general, versus drawing the line at explicitly moral language.

Even though we make no informed hypotheses with respect to the potential links between moral concerns and non-moral areas of language, generating such a set of observations presents an opportunity for an "extensive examination and collection of relevant phenomena and the description of universal or contingent invariances" (Rozin, 2001, p. 3) in the moral domain, which is a vital but underappreciated component of theory development in moral psychology (see Muthukrishna & Henrich, 2019).

These untested assumptions — that moral rhetoric is associated with a congruent set of individual-level moral concerns, and that the connection between moral concerns and language is exclusively through moral rhetoric — motivate the present research, which measures the individual-level associations between observed language on social media and moral concerns, as conceptualized via MFT.

1.1. Moral Foundations Theory

The majority of the studies at the intersection of moral psychology and language (e.g., Araque, Gatti, & Kalimeri, 2020; Mokhberian, Abeliuk, Cummings, & Lerman, 2020; Mooijman et al., 2018; Rezapour, Shah, & Diesner, 2019) rely on MFT as a guiding framework. MFT provides a predictive and pluralistic view of moral concerns that can facilitate an exploration into how such concerns are manifested in language. MFT was developed in order to fill the need of a systematic theory of morality, explaining its evolutionary origins, developmental aspects, and cultural variations. MFT can be viewed as an attempt to specify the psychological mechanisms which allow for intuitive bases of moral judgments as well as moral reasoning. Care, Fairness, Loyalty, Authority, and Purity, according to MFT, are five "foundations" that are conceptualized to have contributed to solving adaptive problems over humans' evolutionary past, and are ubiquitous in current human populations (Graham et al., 2013).

Each of the five foundations in MFT is conceptualized as having solved different adaptive problems in humans' evolutionary past. The Care foundation accounts for our nurturing of the young and caring for the infirm. The Fairness foundation accounts for the development of human cooperation (Purzycki et al., 2018), justice, and reciprocity. Care and Fairness together are considered 'Individualizing' foundations given their emphasis on the well-being and success of individuals. In contrast, the 'Binding' foundations — Loyalty, Authority, and Purity — account for the evolutionarily developed human pursuits of social hierarchy, order, and inherent sanctity or holiness (Graham et al., 2011).

The body of work devoted to studying language within MFT has been motivated by a theoretically-informed text-analytic tool, the Moral Foundations Dictionary (MFD; Graham, Haidt, & Nosek, 2009) which consists of 295 words (and word stems) related to each of the five moral foundations. Words like "peace", "compassion", and "security", when observed in language, are taken to indicate the speaker's endorsement or attention to a particular moral concern (in this case, the Care concern). The taxonomy of MFT — five foundations, and "vice" and "virtue" dimensions of each (e.g., "justice" and "fair" for Fairness virtue, "injustice" and "segregate" for Fairness vice) — has additionally been used to inform annotation of moral sentiment in text (Hoover et al., 2020). MFD has been used in analyzing moral rhetoric and policy debates in social media (e.g., Garten et al., 2018; Mokhberian et al., 2020) as well as in the language of political elites (e.g., Wang & Inbar, 2021).

Two key limitations of MFD-inspired research, which closely align with the two untested assumptions of previous analyses of moral language, further motivate and inform the present work. First, the MFD, and furthermore the categories of moral sentiment proposed by Hoover et al. (2020), have unknown psychometric qualities with respect to moral concerns at the individual-level of analysis. As alluded to above in the discussion of moral language studies, it is unknown whether individual-level differences in usage of MFD words are associated with individual-level differences in moral concerns. For example, it is unknown whether using Purity words is associated with an underlying concern with Purity. Second, previous research at the intersection of morality and language is guided almost exclusively by the MFD, and by the notion of moral language. Exploratory analysis using other categories of language can open up the possibility for new insights from language to be generated for understanding moral concerns, such as affective or social categories of words (Tausczik & Pennebaker, 2010).

1.2. Psychological insight from facebook language

Recently, the potential of using digital records of human behavior from online social media has been realized in many psychological research domains. Examples include the discovery that personality dimensions have been found in Facebook status updates (Garcia & Sikström, 2014; Park et al., 2015; Schwartz et al., 2013), county-level heart disease mortality in the U.S. has been shown to be predicted by Twitter language (Eichstaedt et al., 2015; Eichstaedt et al., 2018), and political elites' usage of moral rhetoric on Twitter was related to their relative levels of power, specifically that U.S. Democrats used more moral language following the election of Republican Donald Trump to the presidency in 2016 (Wang & Inbar, 2021).

In the present work, Facebook status updates were selected for

analysis as a rich, untapped source of insight into the moral domain. Facebook, the most widely-used online social medium in the United States (Facebook, 2019), has been argued to have incredible potential for observational research on individuals in natural contexts (Eichstaedt et al., 2018). Though Facebook language is not "everyday" in the sense that transcribed language recorded by instruments such as the Electronically Activated Recorder (EAR; Mehl, Pennebaker, Crow, Dabbs, & Price, 2001) are, it gives an unobtrusive record of individuals' communication with others and their broadcasting of experiences, emotions, and opinions (Garcia & Sikström, 2014). Eichstaedt, Smith, et al. (2018) give guidelines — both technological and ethical — for the appropriate conduct of Facebook-based research in the social sciences, and we adopt their approach in order to gather naturally-occurring language data from Facebook.

1.3. Overview of the present work

The present work provides the first direct attempt to link individual-level moral concerns and unobtrusively observed language, conducting an investigation of specifically moral language as well as an exploration of language in general. To provide this link, a large sample ($n = 2691$) of online participants' responses to the MFQ (Graham et al., 2008), in addition to their volunteered Facebook status updates ($n = 107{,}798$), was collected.

Analysis 1 tests whether individual-level moral concerns can be predicted from language. This Analysis is a proof-of-concept in two dimensions: first, using dictionary-based techniques, the previously tacitly assumed link between moral language (i.e., measured via MFD) and individual-level moral concerns is tested; second, general measures of language, facilitated by leading NLP techniques, are similarly used to explore the association between language and moral concerns. Analysis 2 is a follow-up exploration of the particular "signatures", or "linguistic traces", of each moral foundation in language. Previous work has used topic models (Park et al., 2015) and dictionaries (Boyd et al., 2015) to identify the correlates of dimensions of personality and personal values, respectively. The present work uses similar tools to build models of language measures as a function of individual-level moral concerns. In what follows, we first present a full description of the data and then Analyses 1 and 2.

2. Data

We recruited a sample of participants who voluntarily completed self-report measures on yourmorals.org and consented to having their Facebook posts accessed for research purposes. This research was reviewed and approved by the University of Southern California's Institutional Review Board (UP-07-00393-AM019).

Status updates were collected, with the approval of Facebook, via the Facebook API in a single bulk retrieval of individuals' posts at the time they completed the survey. Initially, 4414 participants completed the survey, volunteered their Facebook information, and had at least one Facebook post. These participants were filtered due to being outside the age range of 18–65 (592 participants), a similar filter as was applied by Park et al. (2015). After cleaning these participants' Facebook posts of hyperlinks, picture links, and "mentions" using regular expressions and tokenizing text using the Natural Language Toolkit (nltk version 3.4.4; Loper & Bird, 2002) in Python (3.6.7), 53,901 of the 165,787 posts were removed that were either too short (less than five tokens) or could not be recognized as English,[1] using the langdetect Python library (version 1.0.7). Lastly, participants with less than 10 Facebook posts (1131 participants) were removed from the study, similarly to Park et al. (2015). All analyses were repeated with 25 status updates as the threshold and with no threshold at all; these results are included in Supplemental Materials.

This filtering process resulted in 2691 participants. Participants' last status update ranged in time from pre-2016 (20 participants), to 2016 ($n = 1294$), to 2017 ($n = 1377$), reflecting the data collection period, which spanned May 2016 to March 2017. Participants posted an average of 40 status updates ($Mdn = 37$, $SD = 21.4$), with an average of 1158 words per participant ($Mdn = 893$, $SD = 982.4$). The participants' collected posts totaled 107,798, averaging 29.0 tokens in length ($Mdn = 16.0$, $SD = 34.8$). In the full sample of 2691, participants self-reported age ($M = 32.8$, $Mdn = 30.0$, $SD = 11.9$) and sex (Male: 1535, 57.0%; Female: 1156, 43.0%). Though we recognize that gender identity is non-binary, participants were only given the opportunity to identify their binary sex (male or female), which is a limitation of the present work. Participants reported political ideology/political party and religious identification, though these were characterized by high numbers of missing values. For the 488 participants with valid religious identification, 390 identified as Christian/Catholic (79.9%), 66 as Agnostic/Atheist (13.5%), 18 as Jewish (3.7%), and 14 as Other. For the 482 participants with valid political ideology/party, 174 identified as Liberal (37.1%), 113 as Conservative (23.4%), 80 as Moderate/Independent (16.6%), 74 as Libertarian (15.4%), and 41 as Other.

All participants completed the 30-item MFQ, which consists of two 15-item sections, Relevance and Judgments. The first section measures the relevance individuals ascribe to each of the foundations, by asking: "When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?" An example might be "Whether or not someone suffered emotionally". Items on the Relevance section are rated along a 6-point Likert-type scale ranging from 0 (*Not at all relevant*) to 5 (*Extremely relevant*). The Judgments section consists of contextualized items that can gauge actual moral judgments related to the five moral foundations. For example, participants judged the acceptability of "Chastity is an important and valuable virtue". Items on the Judgments section are rated along a 6-point Likert-type scale ranging from 0 (*Strongly disagree*) to 5 (*Strongly agree*). In total, 6 items (3 each for Relevance and Judgment) were collected per foundation, where the above examples correspond to Care and Purity, respectively. The internal consistency coefficients in the present sample were .70, .64, .72, .78, .87 for Care, Fairness, Loyalty, Authority, and Purity.

3. Analysis 1: predicting moral concerns from language

Despite the growing use of moral language (i.e., moral rhetoric or moral sentiment) in studies of social media posts and political persuasion, it is unknown whether moral language and individual-level moral concerns have any meaningful association. Therefore, in this Analysis we test whether individuals' moral concerns can be predicted from measures of their usage of explicit moral language. In addition, the potential relationship between individual-level moral concerns and language extends beyond purely moral language, motivating a general investigation of how language can be used to predict moral concerns. For both moral language and general language, predicting moral concerns from the respective language-based measures is used as an indicator of relatedness, which presents stronger evidence than merely showing a correlation (Yarkoni & Westfall, 2017).

To measure moral language, we rely on dictionaries of moral words, which provide a priori measures of explicit, word-level indicators of target constructs (Pennebaker, Francis, & Booth, 2001) that have been widely used in previous works to measure moral language (e.g., Araque et al., 2020; Dehghani et al., 2016). There are known limitations of this approach, however. The MFD was originally used to find that U.S. religious sermons delivered by liberals and conservatives differed in the amount of moral language used (Graham et al., 2009), but this initial finding was only partially supported in a multi-study replication effort (Frimer, 2020), casting doubt on the validity of the MFD given its

[1] This is necessary given that the dictionaries and pre-processing tools that we use assume English text.

instability across studies. Additionally, the coverage of the MFD has been questioned, as it contains relatively few unique words and stems (approximately 32 per category). As such, MFD measures can be sparse and are prone to missing moral words outside its a priori lexicon. To account for this limitation, the MFD2 (204 words per category) was developed (Frimer, Boghrati, Haidt, Graham, & Dehghani, 2019), and is used in the present analysis in conjunction with the original MFD. However, we note that measuring morality in language is a fundamentally challenging task, made so by its complex linguistic nature — for example, individuals can say moral things without using explicitly moral words. Previous research has yet to directly measure the external validity of MFD-based measures with respect to individual-level moral concerns, which is addressed by the present work.

Beyond the strictly moral domain, our analysis explores other, more general ways that moral concerns are observed in language. To measure language in general, we rely on a variety of techniques, including the established lexicon-based approach of the Linguistic Inquiry and Word Count dictionaries (LIWC; Pennebaker et al., 2001) as well as statistical methods from NLP. These more advanced techniques are based on methodological progress in recent years that allows more accurate measurement of language. Each of the 5 MFQ scores was regressed, in turn, on each of these language representations, with regularization and cross-validation.

3.1. Text representation methods

3.1.1. Dictionaries

Three dictionaries were used in this analysis: the MFD, MFD2, and LIWC's dictionary. Both the MFD and MFD2 have ten categories: the "virtue" and "vice" words for each foundation (indicating opposite polarities for each). MFD additionally has a "General Morality" category (e.g., "wrong", "evil", "good"). The MFD was created by Graham et al. (2009) by first compiling lists of prototypical words, along with synonyms, related terms, and antonyms, associated with each category and refining these lists by removing unrelated words, while the expanded MFD2 was created by Frimer et al. (2019)[2] with the aid of word embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Both the MFD and the MFD2 are domain-specific dictionaries that measure a narrow semantic range of words, and are used here to measure the predictiveness of moral language with respect to individual-level moral concerns. As a more general, a priori measure of language, the default dictionary in the Linguistic Inquiry and Word Count 2015 (LIWC; Pennebaker, Boyd, Jordan, & Blackburn, 2015) was used. The LIWC dictionary contains an exhaustive set of word categories ranging from function words (e.g., personal pronouns, prepositions), grammar (e.g., verbs, past tense), and psychological processes (e.g., cognitive processes, emotion words, social words). In all, 73 lower-level categories are in the hierarchical taxonomy of LIWC (i.e., lower-level categories such as "positive emotions" are nested within "Affective Processes"). Each category in MFD, MFD2, and LIWC was applied in the standard way (Pennebaker et al., 2001), producing a rate of occurrence per category corresponding to the total number of occurrences of a word in a category divided by the total number of words contained in a given participant's posts (see Table 1).

3.1.2. Bag of words modeling

Topic modeling is an effective means of constructing data-driven text representations from word count statistics, where representations are guided by word co-occurrence statistics rather than a priori categories. Most prominently, Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) models documents as mixtures of topics, which are groups of related words that frequently co-occur in documents. For this analysis, LDA was used for its ability to effectively extract relevant information from text, as opposed to its interpretive value for exploring and visualizing trends in a corpus (see Analysis 2).

Table 1
Summary of methods used to extract features for regularized text regression. Word count and Distributed Dictionary Representation (DDR) were used to apply the Moral Foundations Dictionary (MFD), the updated MFD, and the 2015 LIWC. Other general methods for encoding text include Topic Modeling, using Latent Dirichlet Allocation (LDA), Global Vectors for Word Representation (GloVe), and contextualized word embeddings via Bidirectional Encoder Representations from Transformer (BERT) language models.

Method	Description
Word count	Dictionary words in each category are counted and normalized (Pennebaker et al., 2001), using MFD, MFD2, and LIWC
DDR	A hybrid between dictionaries and word embeddings using geometric similarity measures (Garten et al., 2018), using MFD, MFD2, and LIWC
LDA	A statistical model learns word clusters ("topics") by leveraging word co-occurrence information in Facebook posts (Blei et al., 2003)
GloVe	The word embeddings of words in a participant's posts are averaged (Pennington et al., 2014)
BERT	Contextualized word embeddings are extracted from the word-level representation of a full, pre-trained language model (Devlin et al., 2019) and subsequently averaged across words

LDA topics were estimated via mallet,[3] specifically using the Python (v. 3.6) wrapper in the gensim (v. 3.8) package. The mallet implementation of LDA implements collapsed Gibbs sampling (Griffiths & Steyvers, 2004) and automates parameter optimization. The number of latent topics, which is a tunable hyperparameter in LDA and other topic models, was separately tuned internally (as opposed to external measures, such as prediction of individual-level moral concerns for held-out data) using a matrix factorization metric (Arun, Suresh, Madhavan, & Murthy, 2010) via the ldatuning (v 1.0.2) package in R (v 3.6.2). From a search grid across 10–600, a final number of topics was determined to be 300 (see Supplemental Materials). After re-estimating the topic model on the entire corpus of Facebook posts, a single vector of length 300 was estimated per participant using the method of Park et al. (2015). For each participant i, a probability score was computed for each topic, estimated given the fit topic model (which gives probabilities of topics given words):

$$p(\text{topic}|i) = \sum_{w} p(\text{topic}|w) \cdot p(w|i) \quad (1)$$

over participants' normalized word-count proportions for each word w.

3.1.3. Word and text embedding

Neural network-based methods are currently the leading paradigm in NLP, achieving break-through success in predictive modeling in recent years (Devlin, Chang, Lee, & Toutanova, 2019; Pennington, Socher, & Manning, 2014; Vaswani et al., 2017). Though the interpretability and usefulness of these methods for inferential purposes (e.g., learning what types of language are associated with moral concerns) is very much an open question, their capacity for capturing the meaning of language data is unsurpassed in most NLP modeling tasks. In the present analysis, two methods were employed, word embeddings and contextualized word embeddings. Word embeddings (Mikolov et al., 2013; Pennington et al., 2014) map words into a geometric space that preserves detailed semantic and syntactic information. For each participant's set of posts, GloVe word embedding vectors (Pennington et al., 2014), which were trained on text from the Common Crawl,[4] were mapped to the individual words in the posts and subsequently averaged element-wise (i.e., across the dimensions of the embedding). The result was a single vector, per participant, which was the "average" word embedding that occurred in

[2] https://osf.io/ezn37/

[3] http://mallet.cs.umass.edu

[4] Data are available at https://nlp.stanford.edu/projects/glove/

their posts.[5]

In addition, we apply Distributed Dictionary Representations (DDR; Garten et al., 2018) to the dictionaries used above (MFD, MFD2, and LIWC). DDR operates by first computing the average (element-wise) of each dictionary, which represent the semantic space occupied collectively by the dictionary category in question. Then, to compute the "loading" of a dictionary on a particular piece of text, the text's average word embedding is similarly computed, and the geometric similarity is then computed between these two vectors.

We also use contextualized word embeddings, which extract word embeddings from pretrained language models (e.g., Devlin et al., 2019; Peters et al., 2018), which are "contextual" because a given word can have multiple embeddings based on context. Whereas word embeddings such as GloVe produce the same embedding for a given word, no matter the context, contextualized embeddings compute the embedding of each word while taking its context into account. The importance of this contextual flexibility is illustrated by considering words like "bank", which can be in reference to a financial institution, the shore of a river, or a verb. In practice, contextual word embeddings are generated by feeding sequences of words into large neural network models, computing these dynamic, contextual word vectors for each word, and taking their average. The information contained in these contextualized embeddings is comparatively richer than word embeddings like GloVe, given their ability to generate representations that are dependent on the composed meaning of words in sequence, rather than each word independent of context.

The present analysis used a previously trained instance of the Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). BERT is a model that processes words *in sequence* through multiple layers of encoding, at each layer capturing more sequentially-dependent information (Vaswani et al., 2017). The BERT model has been trained, using large datasets, on text comprehension tasks, such as predicting masked-out words and whether one sentence follows another. These models are available for download, allowing researchers to generate contextualized word embeddings by passing segments of text through the previously-learned layers of BERT. BERT vectors were produced for a given sequence of tokens using the transformers[6] library in the Python (v3.6) programming language, where the last four layers (of a possible 12) were averaged to form a single vector of length 768. During experimentation, it was found that the best predictability via BERT was achieved by computing embedding vectors for each post and subsequently averaging them, rather than computing a single embedding across all of a participants' words (see Supplemental Materials).

Each method used for generating text representations is summarized in Table 1, which first lists dictionary-based methods (word count, DDR), which were applied to each of the three dictionaries, as well as three general techniques (LDA, GloVe, BERT).

3.2. Regression analysis

Every representation produced from the above methods was uniformly used as input in an ElasticNet regression for predicting each MFQ score. ElasticNet allows tunable regularization, combining both LASSO and Ridge regression penalties (Zou & Hastie, 2005) to reduce model complexity and enable optimal feature selection, which is desired for high-dimensional data. Models were implemented in Python (v 3.6) using the scikit-learn library (v 0.22 Pedregosa et al., 2011), with the ratio between L1 and L2 penalties and the overall regularization term determined through maximizing R^2 with cross validated grid search. To obtain estimates of variance for test-set R^2s, 5-fold cross-validation was repeated 10 times. ElasticNet parameters were selected within each fold's training data and were not tuned to the test data for that fold.

3.3. Results

Full statistics on model performance across 10 rounds of 5-fold cross validation for each text representation method for each foundation are provided in Table 2. Bold-face values indicate the highest R^2 achieved within +/− one SE of each other, per foundation and method categorization.

For each foundation, a repeated measures Analysis of Variance (ANOVA) was performed, with representation method as the grouping variable. Nine possible groups were compared, consisting of the nine representation methods used. R^2 values from the 10 repeated 5-fold cross-validation runs were modeled as the dependent variable. Each set of values satisfied the Shapiro-Wilk test of normality and Levene's test for homogeneity of variance between groups. There was a significant overall effect of representation for predicting Care ($F(8, 441) = 89.0$, $p < 0.001$), Fairness ($F(8, 441) = 54.4$, $p < 0.001$), Loyalty ($F(8, 441) = 117.3$, $p < 0.001$), Authority ($F(8, 441) = 327.5$, $p < 0.001$), and Purity concerns ($F(8, 441) = 299.9$, $p < 0.001$).

It is clear from Table 2 that MFD-based measures are in general less effective at predicting individual-level moral concerns than are LIWC, LDA, GloVe, and BERT methods. In other words, the frequency with which participants used explicitly moral words, such as "sacred", "loyal", or "compassion", was only slightly related to moral concerns when compared to measures of general language, such as a participant's average word embedding. Multiple operationalizations of moral dictionaries were tested in this analysis, which showed that the most effective MFD-based measure also varied by foundation: for predicting Care concerns, MFD_{DDR} was significantly more effective than both word counting methods and $MFD2_{DDR}$, but MFD_{DDR} and $MFD2_{DDR}$ were not significantly different in predicting other concerns. This finding can be investigated in future work, as it may indicate the MFD's differences, among foundations, in coverage of relevant words.

Tukey-adjusted post-hoc pairwise comparisons were performed to assess the effects of more general text representation methods. For Care, $LIWC_{DDR}$ and LDA representations were not significantly different from each other while both explained more variance than LIWC ($ps < 0.001$). For Loyalty and Purity, all three methods were not significantly different from each other, whereas LDA representations explained more variance than both LIWC and $LIWC_{DDR}$ for Authority and $LIWC_{DDR}$ explained

Table 2
Percent variance explained (R^2) across representation methods per foundation. Bold values are highest within category (measures of moral language; general measures) for given foundation.

Representation	Care	Fairness	Loyalty	Authority	Purity
Measures of moral language					
MFD	1.0 (0.1)	0.3 (0.1)	1.3 (0.2)	1.6 (0.2)	2.2 (0.2)
MFD2	2.7 (0.2)	0.6 (0.1)	2.8 (0.2)	4.8 (0.2)	**10.0 (0.3)**
MFD_{DDR}	4.8 (0.3)	0.9 (0.1)	4.9 (0.2)	7.2 (0.3)	8.2 (0.3)
$MFD2_{DDR}$	3.0 (0.2)	−0.3 (0.1)	6.1 (0.3)	8.0 (0.3)	**10.1 (0.4)**
General measures of language					
LIWC	4.0 (0.2)	0.7 (0.1)	8.3 (0.3)	13.7 (0.4)	16.8 (0.4)
$LIWC_{DDR}$	5.7 (0.3)	2.4 (0.2)	8.0 (0.5)	13.7 (0.5)	16.5 (0.4)
LDA	5.9 (0.3)	1.2 (0.2)	8.8 (0.3)	15.6 (0.3)	17.2 (0.3)
GloVe	7.2 (0.3)	2.9 (0.2)	10.9 (0.4)	17.7 (0.4)	**20.0 (0.5)**
BERT	8.8 (0.3)	3.2 (0.2)	11.7 (0.4)	18.8 (0.3)	20.9 (0.4)

Note. Mean and standard errors across 10 iterations of 5-fold cross validation ($n = 50$). If two values' standard errors overlap and are highest, both are displayed in bold.

[5] Alternatively, average embeddings can be generated per post and subsequently averaged to a single participant-level vector, but this was found to lower predictive results (see Supplemental Materials)

[6] https://huggingface.co/transformers/

more variance than LIWC for Fairness.

For all concerns, GloVe and BERT vectors explained more variance than all other methods ($ps < 0.05$) with the exception that $LIWC_{DDR}$ and GloVe were not significantly different for Fairness concerns ($p = 0.47$). GloVe and BERT yielded similar levels of information across foundations, where only for Care did BERT explain significantly more variance than GloVe ($p = 0.003$). Overall, the higher levels of explained variance for GloVe and BERT embeddings shows that there are differences in language with respect to moral concerns that are not captured by individuals' lexicons (i.e., LIWC and LDA). Intuitively, individuals with different moral concerns occupy different "semantic spaces", which are captured by embeddings at a more fine-grained level than LIWC or LDA, while the greatest "distances" in this space were observed for Purity and Authority.

To quantify the differences between foundations in overall predictability, an additional two-way repeated measures ANOVA assessing the influence of representation and foundation on explained variance was performed. Specifically, the influences of representation (9 groups) and foundation (5 groups) on average R2 was measured, as well as their interaction. There was a main effect of representation ($F(8, 2205) = 812.2, p < 0.001$) and foundation ($F(4, 2205) = 2475.6, p < 0.001$), with a significant interaction between the two ($F(32, 2205) = 67.4$, $p < 0.001$). Tukey-adjusted post-hoc pairwise contrasts showed that differences among foundations were uniformly significant at $p < 0.001$; in particular, R^2 values for Purity were higher than all other foundations, whereas R^2 values were lowest for Fairness. Visualizations of the full set of comparisons for this two-way ANOVA are shown in the Supplemental Materials.

3.4. Discussion

There is indeed a relationship between social media language and moral concerns as measured by the MFQ, which varies significantly among foundations. Though the magnitude of this relationship changes in terms of how text data is quantified, it is apparent from our analysis that moral concerns coexist with different patterns of expression and communication on social media. Measures of moral language, specifically via the MFD and its variants, predicted individual-level concerns to marginal degrees, especially when compared to general-purpose measures. In particular, only about 1% of Fairness concerns' variance could be explained by MFD-based measures.

The range of methodologies used, specifically their difference in explained variance, shed light on what drives the link between moral concerns and language. LIWC and LDA, which yield similar predictiveness to each other, are able to pick up on individuals' *lexical* differences — the categories of words and topics they mention, as well as whether they spend more or less time speaking with a certain style (e.g., using personal pronouns). In contrast, MFD-based representations describe individuals' usage of specifically moral words. Given that LIWC and LDA representations described significantly more variance in individuals' moral concerns, the lexical differences tied to moral concerns extend far beyond the strictly moral. This is explored further in Analysis 2.

Further, we found that embedding methods consistently explained more variance than LIWC and LDA. Since embedding methods are superior predictive approaches and are known to contain essential linguistic information, they likely provide the most accurate indicator of the relationship between moral concerns and language. One possible explanation for the higher predictiveness of embeddings is that individuals' linguistic differences, in relation to their moral concerns, are captured by GloVe and BERT at a more fine-grained level. GloVe embeddings can capture the fact that not only are "hurt", "ugly", and "nasty" in the same word category ("negative emotion" in LIWC; Tausczik & Pennebaker, 2010, Appendix), but also that *within* this category, "nasty" and "ugly" are more similar in meaning than "nasty" is to "hurt" or than "ugly" is to "hurt". In addition, contextualized BERT embeddings are further able to measure linguistic differences by accounting for words in the context of their surroundings.

Though representations predict moral concerns to varying degrees, the differences between foundations are robust with respect to representation technique. Individuals' Purity concerns are strongly tied to language, in such a way that even relatively naive quantification techniques can achieve high explanation coefficients, while sophisticated techniques like BERT are able to achieve even higher explanation coefficients. Fairness concerns had the weakest association with language, regardless of the method employed.

4. Analysis 2: signatures of moral concerns in language

In Analysis 2, we explore the signatures of each moral concern in language using the interpretable NLP techniques from Analysis 1. Specifically, we use dictionaries and topic models, rather than neural network-based embedding approaches that might be superior predictors but are more opaque. Dictionaries yield interpretable measures from text that are based on face-valid categorizations of words, with transparent construction and wide usage contextualizing downstream inferences. Topic models like LDA are particularly appropriate for visualizing and exploring text corpora (e.g., Eichstaedt et al., 2015), and are complementary to dictionaries in their ability to uncover word categories that are particular to a given dataset, and not captured by predefined dictionary categories.

In this analysis, language-based measures were modeled as functionally dependent on individual-level moral concerns. In other words, we test whether variance in each dimension of language usage can be explained collectively by individuals' moral concerns. Unique "signatures" of each moral concern are more easily identifiable in such models, given the default post-hoc interpretation of regression analysis — i.e., the effect of moral concerns on language can be examined, per concern, while keeping other individual-level variables constant. This is in contrast to previous work analyzing the relationship between social media language and personality traits, which largely regress participants' traits on text representations (e.g., Park et al., 2015; H. A. Schwartz et al., 2013) resulting in models that do not allow direct interpretations of the particular effect of each moral concern on language-based outcomes.

4.1. Method

Three sets of linguistic features — MFD2, LIWC, and LDA topic model features — were modeled as dependent variables in separate regressions. The MFD2 and LIWC dictionaries were applied in the same way as in Analysis 1, but without normalizing raw word counts (as models used were offset regressions of counts). Word-counts were chosen given that counts provide intuitive effects (i.e., rate of usage versus a single similarity score). The LDA text representations computed in (1) using the same model as Analysis 1 were used as topic probabilities for each participant.

Each word count outcome (for MFD2 and LIWC) was modeled as a dependent variable in a separate negative binomial regression with offsets for the logarithm of total word count, accounting for the fact that participants had varying numbers of words in their posts. Thus, coefficients of these offset negative binomial models correspond to the change in the rate of the particular category being modeled. LDA topic outcomes were modeled using linear regressions. For each outcome, two models were fit: one with foundation-level MFQ scores (standardized) as independent variables, and one that included MFQ scores as well as controlled for the two demographic variables that were available in the full dataset, age and sex. All models were applied to the same set of observations ($N = 2691$), namely, individual-level measurements of language, MFQ, and demographics. Negative binomial models were fit in R (Version 3.6; R Core Team, 2019) using the MASS package (Version 7.3–51.5; Ripley et al., 2013).

4.2. Results

Results are presented first for MFD2, which address questions as to the correspondence between moral language and moral concerns. Specifically, we determine the categories of moral language that are predicted distinctly by each moral concern. Secondly, exploratory results are shown for the LIWC dictionary categories and for LDA topics.

4.2.1. Moral language

Fig. 1 shows β values for negative binomial models of MFD2 outcomes with offsets for total number of words by participants. p-values were corrected for multiple comparisons. Specifically, for models without demographics, bonferroni corrections were made for 5 (MFQ) predictors, and for models with demographics, corrections were made for 7 predictors. Of primary interest is whether higher moral concerns predict higher rates of moral language, in the corresponding moral dimension (e.g., higher Care concerns predicting more Care language). This was found to be the case for Purity ($\beta = 0.350$, $SE = 0.026$, $p < 0.001$), Care ($\beta = 0.166$, $SE = 0.015$, $p < 0.001$), and Fairness ($\beta = 0.106$, $SE = 0.031$, $p = 0.003$), in the model without participant age and sex. Loyalty ($\beta = -0.018$, $SE = 0.020$, $p = 1.0$) and Authority ($\beta = -0.039$, $SE = 0.026$, $p = 1.0$) did not predict differences in the corresponding category of moral language. Dimensions of *vice* (versus *virtue*) within each category of moral language were, for the most part, not associated with higher moral concerns. One exception was Cheating (Fairness) moral language, with a 1 SD increase predicting significantly more Cheating language ($\beta = 0.113$, $SE = 0.036$, $p = 0.009$).

Outside of the relationships between moral concerns and their corresponding dimension of moral language, Purity and Fairness concerns both predicted differences in several categories of moral language. Higher Purity concerns predicted a lower rate of Degradation language ($\beta = -0.132$, $SE = 0.032$, $p < 0.001$) and a higher rate of Betrayal language ($\beta = 0.220$, $SE = 0.088$, $p = 0.006$). Higher Loyalty concerns predicted a lower rate of Betrayal language ($\beta = -0.167$, $SE = 0.088$, $p = 0.019$), and higher Fairness concerns predicted higher rates of Degradation ($\beta = 0.082$, $SE = 0.035$, $p = 0.02$) and Harm ($\beta = 0.078$, $SE = 0.030$, $p = 0.009$).

In terms of participant age and sex, women posted a significantly higher rate of Care language ($\beta = -0.380$, $SE = 0.025$, $p < 0.001$), while the effect of Care concerns on Care language was still significant after controlling for age and sex ($\beta = 0.104$, $SE = 0.015$, $p < 0.001$). Men used significantly more words in the Cheating ($\beta = 0.421$, $SE = 0.062$, $p < 0.001$), Fairness ($\beta = 0.257$, $SE = 0.053$, $p < 0.001$), Betrayal ($\beta = 0.973$, $SE = 0.138$, $p < 0.001$), and Authority ($\beta = 0.308$, $SE = 0.033$, $p < 0.001$) categories. Lastly, higher rates of Authority language ($\beta = 0.106$, $SE = 0.015$, $p < 0.001$) and Loyalty language ($\beta = 0.058$, $SE = 0.013$, $p < 0.001$) were predicted with 1 SD increases in participant age.

4.2.2. LIWC categories

Fig. 2 shows the results from offset negative binomial models of LIWC word counts. Higher Purity, which our first analysis showed to be the most traceable concern, predicts significantly less sexual language ($\beta = -0.193$, $SE = 0.043$, $p < 0.001$) and swearing ($\beta = -0.172$, $SE = 0.038$, $p < 0.001$). These effects were robust to inclusion of age and sex in the models (see Fig. 2). Additionally, higher Care is associated with an increase in social language (i.e., personal pronouns, friends, family, affiliation), positive emotion, and health-oriented language, though many of these relationships do not persist in models adjusting for age and sex. Fairness and Loyalty concerns did not predict any coherent set of word topics, when adjusting for other concerns and participant age and sex.

4.2.3. Topic categories

Here we report results for the LDA models that controlled for participant age and sex. For each MFQ predictor, the topics which were predicted with the highest positive value (i.e., a positive association between MFQ predictor and topic outcome) are reported in Fig. 3. Topics that were not significant ($p < 0.01$) were not reported. Topics are displayed using the top 8 terms, and coefficients are interpreted as the percent-increase in the probability of a given topic occurring in a participant's posts given a 1 SD increase in the predictor, adjusting for other predictors.

A number of patterns emerge in Fig. 3 which were not shown apparent from LIWC or MFD2 analyses. Higher Fairness concerns predict higher attention to political topics, for example healthcare, human rights, and gun control, whereas higher Loyalty concerns predict attention to team sports and honoring military service members.

To inform the interpretation of the topics visualized in Fig. 3, three metrics proposed by Mimno, Wallach, Talley, Leenders, and McCallum (2011) for evaluating topic models — coherence, document entropy, and exclusivity — are reported in the Supplemental Materials. These three metrics, respectively, refer to the coherence of individual topics, whether a topic is concentrated over few documents (low entropy), and how distinct an individual topic is versus other topics. Of note, topics included under Authority were, on average, lower in document entropy, highly coherent, and highly exclusive. Topics for Purity were higher in document entropy and lower in exclusivity, and the lowest levels of coherence were observed for Loyalty.

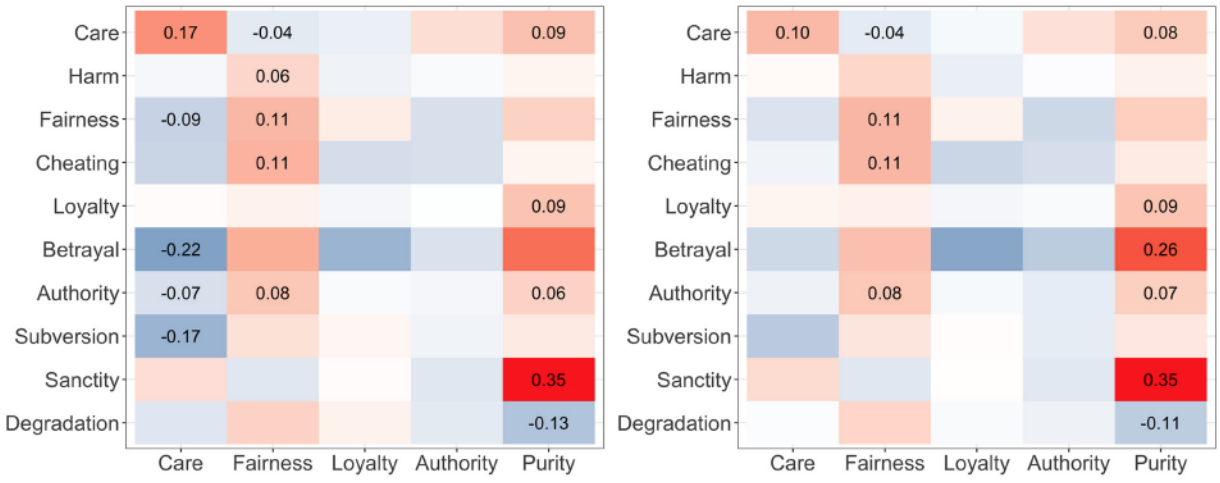

Fig. 1. Coefficients for models of each category (rows) in the updated Moral Foundations Dictionary (MFD2), with offsets for participants' total word count. Values indicate the expected rate increase in the MFD2 outcome, given a 1 SD increase in the Moral Foundations Questionnaire (MFQ) predictor. Cells without numbered coefficients were not significant ($p < 0.05$, bonferroni-corrected for comparisons across number of predictors).

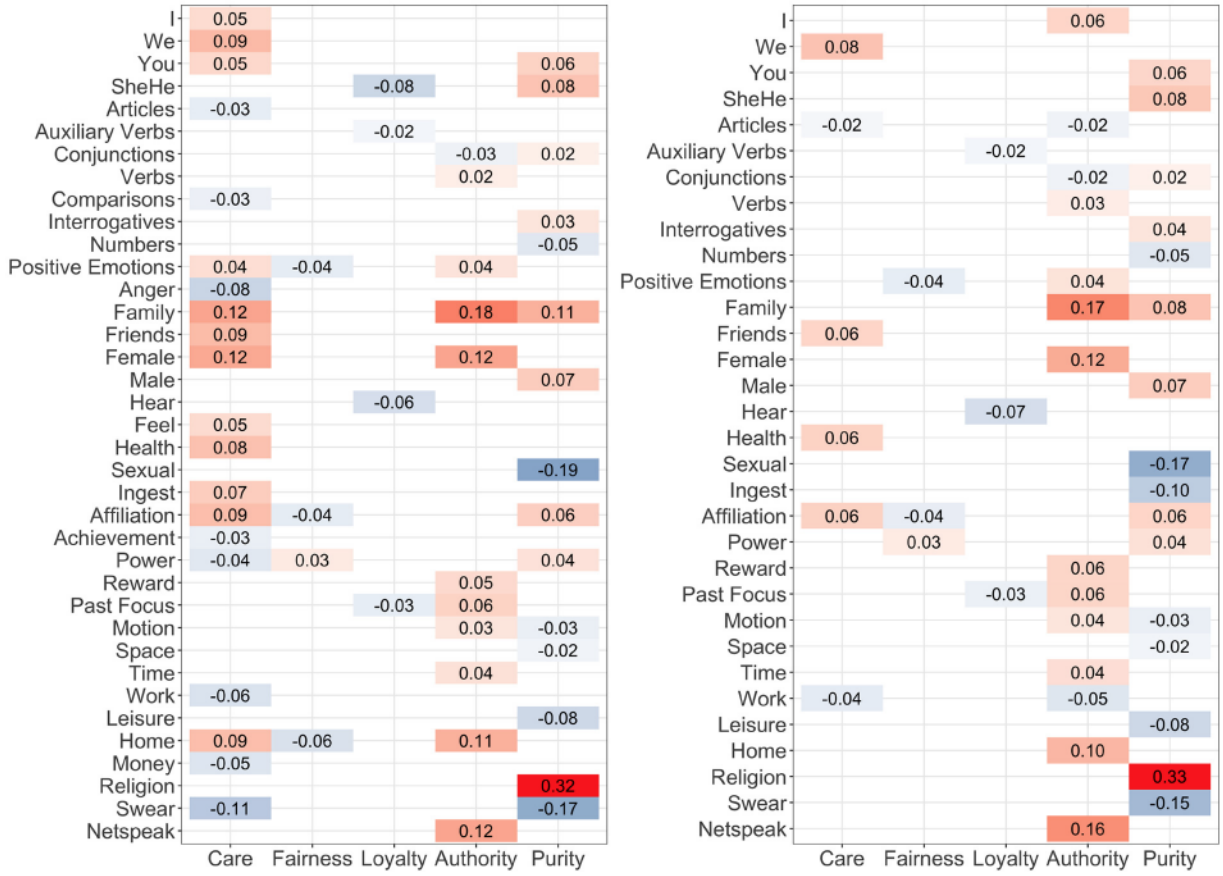

Fig. 2. Coefficients for negative binomial models of each category (rows) in the Linguistic Inquiry and Word Count (LIWC) 2015 release, with offsets for participants' total word count. Values indicate the expected rate increase in the LIWC outcome given a 1 SD increase in the Moral Foundations Questionnaire (MFQ) predictor. Cells without numbered coefficients were not significant (bonferroni-corrected for comparisons across number of predictors), and rows without any significant MFQ predictors were dropped.

4.3. Discussion

In this analysis we determined the distinct signatures of each moral concern in language, using a combination of pre-defined dictionaries and topic modeling. By controlling for demographics and modeling language-based outcomes as dependent on moral concerns, we generated interpretable measures of the extent to which each moral concern influenced the occurrence of a given category of language. We first found that Care, Fairness, and Purity concerns were positively associated with the usage of words from the corresponding MFD2 categories, and used LIWC and LDA-based measures to explore other facets of language: Care concerns specifically predict charity- and health-related posts, Fairness concerns predict political and justice-related language, and Purity concerns motivate prayer, biblical quotation, and gratitude (toward God). Though effects between moral concern and moral language were not found for Authority and Loyalty, associations were found between Authority and language about family, socializing, and daily life, and between Loyalty and language about military members and team sports.

5. General discussion

This exploratory study generated new insight into the relationship between individual-level moral concerns and language, both in terms of moral language, which has been extensively used by previous studies to analyze morality in ecologically valid settings, and in other, more general facets of language use. In Analysis 1, all moral concerns except Fairness were predicted from general language measures with sizeable effect sizes, with the highest for Purity. For all moral concerns, moral language measures were able to explain less variance than general measures. In Analysis 2, associations between moral and non-moral categories of language and moral concerns were found, including an association between religious language and Purity, between social language and Authority, Care, and Purity, and between political topics and Fairness.

While associating observed language with validated measures is not new in psychology, the present work is the first attempt to do so in the moral domain. While moral language has been observed in naturalistic contexts (e.g. Mokhberian et al., 2020; Mooijman et al., 2018), these measures are not anchored at the individual level to psychometrically validated measures. To ensure the validity and consistency of evidence generated for this new domain, we expanded on the methodologies of previous work (Park et al., 2015; Schwartz et al., 2013), in two ways: (a) a diverse collection of NLP methods to contextualize effect sizes, and (b) modeling language-based outcomes as dependent variables in order to isolate relationships between individual-level moral concerns and language.

The findings of Analysis 1 indicate that each moral concern can be differentially predicted from language. The highest explained variance was observed for Purity, followed by Authority. Indeed, Purity and Authority concerns have been previously shown to be predictive of ideological disagreements and culture wars (Graham et al., 2009; Koleva, Graham, Iyer, Ditto, & Haidt, 2012), and thus might be expected to manifest in markedly different categories of social language. Care and Fairness, which have lower variance overall and do not have the same predictiveness of political ideology as do the Binding foundations, are

Fig. 3. Moral Foundations Questionnaire (MFQ) coefficients of linear models predicting each participant-level topic probability, generated via Latent Dirichlet Allocation (LDA). Printed coefficients were multiplied by 100, thus are at percent level. Each row is a model, and all models included age and gender.

predictable but only to a limited extent. Fairness in particular was found to be least traceable in language among the five moral foundations.

In Analysis 2, the types of language associated with moral concerns were markedly different, even where explained variance was similar. In all five cases, there was at least one category of language (either dictionary category or LDA topic) which could be intuitively associated with preconceptions of the given foundation: Care concerns predicted Care virtue language, specifically familial words and positive affective language; Fairness predicted Fairness virtue and vice language, specifically posts talking about general concerns for members of society, such as health care, protests, and climate change; Loyalty predicted language about team sports; Authority predicted familial language and accounts of social experiences; and Purity predicted a large array of religious and spiritual categories of language, combining biblical quotes with publicly shared prayer. These findings are indicative of the fact that moral concerns do manifest in public displays and endorsements which are particular to each moral concern, to varying degrees.

It is notable that Authority — and to a lesser extent, Loyalty — are positively associated with an array of language usage patterns, but these are not captured by the MFD. This indicates that Loyalty and Authority concerns do not manifest in language through direct endorsement or recognition of the corresponding virtues (as is true for Care, Fairness, and Purity). Instead, they manifest through higher rates of references to family and friends, a higher rate of personal pronouns, more "netspeak" (i.e., colloquial language), and more sharing about team sports. Though it is more apparent in the MFD-based measures, the same is also true for Care. Higher Care values predict significantly more "we" pronouns and posts about personal health and affiliation, which are not explicitly modeled in the MFD operationalization of moral language. Those high in Fairness, on the other hand, were preoccupied with political topics and violations of equality or justice; among all concerns, it is the only one for which a positive effect is seen for "vice" language in the corresponding moral category. These findings reflect on the many ways that individual-level moral concerns manifest in language, and that the current view of moral language — in particular, measured via the MFD — might not be accurately capturing these signals.

5.1. Limitations

A number of limitations condition the interpretation of our findings, while simultaneously identifying directions for future work. Internet-based questionnaire responses, such as the MFQ responses gathered in the present study, have great potential for expanding the scale and diversity of study samples (Gosling, Vazire, Srivastava, & John, 2004), but it is known that they are not representative across cultures. Participants completing the MFQ on yourmorals.org do so voluntarily, as do those who grant access to their Facebook accounts. This self-selection bias poses potential confounds for our study, and encourages the collection of similar data through other venues or by collecting more meta-data on participants.

Participants of the study were predominately from Western, Educated, Industrialized, Rich, and Democratic societies (WEIRD; Henrich, Heine, & Norenzayan, 2010, also see Atari, Graham, & Dehghani, 2020 for a discussion on non-WEIRD moral psychology). It is unknown whether the findings in this study would replicate in non-WEIRD populations, which constitute the vast majority of the world's population. Since MFT was developed, in part, to establish a descriptive account of moral concerns outside those observed in the secular West, the generalization of the present findings outside WEIRD samples is important for understanding the general relationship between moral foundations and language.

Additionally, the English-centricity of NLP has been the focus of much recent criticism and emphasis (e.g., Bender & Friedman, 2018), and much of the resources for text analysis in psychology are exclusively in English or are translated from English (e.g., the Japanese MFD; Matsuo, Sasahara, Taguchi, & Karasawa, 2019). These findings cannot be generalized to non-English-speaking cultures. Though we are unable to extend the present work toward other languages, due to the available sample population and the language of dictionaries we use, we acknowledge that the present findings do not necessarily generalize to other languages. Bridging the language gap can be one way in which the present work is replicated and expanded in other cultures, as this would require both non-WEIRD research participants and non-WEIRD researchers (i.e., researchers fluent in non-English languages and non-WEIRD morality Medin, Ojalehto, Marin, & Bang, 2017).

Lastly, our work builds upon the theoretical framework provided by MFT in compiling evidence about individuals' moral concerns. Of course, there are other theories of the fundamental structure of human morality, with some arguing for these alternative theories over MFT. In particular, other theories address the structure of humans' underlying values (S. H. Schwartz, 1992), the dyadic structure of morality in terms of "moral agents" and "moral patients" (Gray & Wegner, 2009), morality as cooperation (Curry, Whitehouse, & Mullins, 2019), relational contexts (Rai & Fiske, 2011), the motivational emphasis of morality (Janoff-Bulman & Carnes, 2013), and the normative emphasis of morality (actions vs. consequences; Cushman, 2013; Mikhail, 2007). The present work is the first study in moral psychology to directly compare social media language and measures of moral concerns from a validated survey; thus, our analyses were restricted to the available data, which only extended to MFT. Future data collection and language analyses can be completed for other theories, allowing comparisons and a richer, more general set of inferences about language and moral concerns.

6. Conclusion

Among the five moral foundations (Care, Fairness, Loyalty, Authority, and Purity), Purity concerns are most traceable in social media language. Fairness concerns, on the other hand, are least traceable. Individuals who highly endorsed Purity shared religious and spiritual content on Facebook, whereas people who scored higher on Fairness were slightly more likely to share content related to social justice and equality. High levels of Care, Loyalty, and Authority were found to motivate a mixed collection of socially-oriented language categories. The link between moral concerns and language was found to extend beyond exclusively moral language. Overall, this research establishes a missing link in moral psychology by providing evidence that individual-level moral concerns are differentially associated with language data collected from individuals' Facebook accounts.

Author note

All code is available on GitHub (https://github.com/BrendanKennedy/moral-concerns-in-language) and all anonymized data have been publicly deposited to the Open Science Framework repository (https://osf.io/jcuqk/). This research was sponsored by NSF CAREER BCS-1846531 to MD.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104696.

References

Araque, O., Gatti, L., & Kalimeri, K. (2020). Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems, 191*, 105184.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 391–402).

Atari, M., Graham, J., & Dehghani, M. (2020). Foundations of morality in Iran. *Evolution and Human Behavior, 41*, 367–384.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. In *Icwsm* (pp. 31–40).

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*, 7313–7318.

Burton, J., Cruz, N., & Hahn, U. (2019). How real is moral contagion in online social networks?. In *Cogsci* (pp. 175–181).

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods, 47*, 1178–1198.

Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *The Journal of Politics, 75*, 659–671.

Curry, O., Whitehouse, H., & Mullins, D. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology, 60*(1).

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin, 40*, 1559–1573.

Dehghani, M., Johnson, K., Hoover, J., Sagi, E., Garten, J., Parmar, N. J., … Graham, J. (2016). Purity homophily in social networks. *Journal of Experimental Psychology: General, 145*, 366.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Eichstaedt, J. C., Schwartz, H. A., Giorgi, S., Kern, M. L., Park, G., Sap, M., … Ungar, L. H. (2018, Mar). More evidence that twitter language predicts heart disease: A response and replication. *PsyArXiv*. https://doi.org/10.31234/osf.io/p75ku. Retrieved from psyarxiv.com/p75ku.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., … Seligman, M. E. P. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science, 26*, 159–169.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preotiuc-Pietro, D., … Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences, 115*(44), 11203–11208.

Facebook. (2019). Number of monthly active facebook users worldwide as of 2nd quarter 2019 (in millions) [graph]. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

Feinberg, M., & Willer, R. (2015). From gulf to bridge: When do moral arguments facilitate political influence? *Personality and Social Psychology Bulletin, 41*(12), 1665–1681.

Frimer, J. (2020). Do liberals and conservatives use different moral languages? Two replications and six extensions of graham, haidt, and nosek's (2009) moral text analysis. *Journal of Research in Personality, 84*, 103906.

Frimer, J., Boghrati, R., Haidt, J., Graham, J., & Dehghani, M. (2019). *Moral foundations dictionary for linguistic analyses 2.0. (unpublished manuscript)*.

Garcia, D., & Sikström, S. (2014). The dark side of facebook: Semantic representations of status updates predict the dark triad of personality. *Personality and Individual Differences, 67*, 92–96.

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods, 50*, 344–361.

Gilligan, C. (1977). In a different voice: Women's conceptions of self and of morality. *Harvard Educational Review, 47*(4), 481–517.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *, Vol. 47. Advances in experimental social psychology* (pp. 55–130). Elsevier.

Graham, J., Haidt, J., & Nosek, B. A. (2008). *The moral foundations questionnaire* (MoralFoundations. org).

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*, 1029–1046.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*, 366–385.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl. 1), 5228–5235.

Grover, T., Bayraktaroglu, E., Mark, G., & Rho, E. H. R. (2019). Moral and affective differences in us immigration policy debate on twitter. *Computer Supported Cooperative Work (CSCW), 28*(3–4), 317–355.

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*, 55–66.

Hallen, B. (2000). *The good, the bad, and the beautiful: Discourse about values in yoruba culture*. Indiana University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–83.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345*(6202), 1340–1343.

Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., … Dehghani, M. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science, 11*(8), 1057–1071, 19485506198766**2**.

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review, 17*(3), 219–236.

Kohlberg, L. (1981). *Essays on moral development: The psychology of moral development* (vol. 2). San Francisco: Harper & row.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. *Journal of Research in Personality, 46*, 184–194.

Li, L., & Tomasello, M. (2021). On the moral functions of language. *Social Cognition, 39*(1), 99–116.

Loper, E., & Bird, S. (2002). *Nltk: The natural language toolkit*. arXiv preprint cs/0205028.

Matsuo, A., Sasahara, K., Taguchi, Y., & Karasawa, M. (2019). Development and validation of the Japanese moral foundations dictionary. *PLoS One, 14*.

Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2017). Systems of (non-) diversity. *Nature Human Behaviour, 1*(5), 1–5.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers, 33*, 517–523.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272).

Mokhberian, N., Abeliuk, A., Cummings, P., & Lerman, K. (2020). Moral framing and ideological bias of news. In S. Aref, et al. (Eds.), *Social informatics* (pp. 206–219).

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour, 2*, 389.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*, 221–229.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*, 934–952.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pennebaker, J. W. (2013). *The secret life of pronouns: What our words say about us*. Bloomsbury Publishing USA.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015 (tech. Rep.)*. Austin, TX: University of Texas at Austin.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: Liwc 2001. 71* p. 2001). Mahway: Lawrence Erlbaum Associates.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547–577.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20*(2), 188–214.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). Association for Computational Linguistics.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Purzycki, B. G., Pisor, A. C., Apicella, C., Atkinson, Q., Cohen, E., Henrich, J., … Xygalatas, D. (2018). The cognitive and cultural foundations of moral behavior. *Evolution and Human Behavior, 39*(5), 490–501. SEP. Doi: {10.1016/j.evolhumbehav.2018.04.004}.

R Core Team. (2019). R: A language and environment for statistical computing [computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*(1), 57.

Rezapour, R., Shah, S. H., & Diesner, J. (2019). Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 35–45).

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package "mass". *Cran R, 538*. Retrieved from http://cran.r-project.org/web/packages/MASS/MASS.pdf.

Rodriguez, A. J., Holleran, S. E., & Mehl, M. R. (2010). Reading between the lines: The lay assessment of subclinical depression from written self-descriptions. *Journal of Personality, 78*(2), 575–598.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2–14.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … others. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS One, 8*, Article e73791.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology, 25*(1), 1–65.

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology, 32*(5), 519–542.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology, 29*, 24–54.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Voelkel, J. G., & Feinberg, M. (2018). Morally reframed arguments can affect support for political candidates. *Social Psychological and Personality Science, 9*(8), 917–924.

Wang, S.-Y. N., & Inbar, Y. (2021). Moral language use by U.S. political elites. *Psychological Science, 32*(1), 14–26.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B: Statistical Methodology, 67*(2), 301–320.